Algorithms for cross-lingual data interlinking

Tatiana Lesnikova, Jérôme David, Jérôme Euzenat

NSFC-61261130588

ANR-12-IS02-002-01

Lindicle

Linked data interlinking in a cross-lingual environment
跨语言环境中语义链接关键技术研究
Liage des données dans un environnement interlingue

D4.2 Algorithms for cross-lingual data interlinking

Coordinator: Tatiana Lesnikova
With contributions from: Tatiana Lesnikova, Jérôme David, Jérôme Euzenat

Quality reviewer:	Jérôme David
Reference:	Lindicle/D4.2/v8
Project:	Lindicle ANR-NSFC Joint project
Date:	July 8, 2015
Version:	8
State:	final
Destination:	public

Executive Summary

Cross-lingual data interlinking consists of discovering links between identical resources across data sets in different languages. The use of different languages renders difficult the comparison of these resources.

We present a general framework for interlinking resources in different languages based on:

- using linguistic elements of the RDF data sets to associate a representation to each resource;
- measuring a similarity between representations;
- extracting links based on the computed similarity.

Three methods have been defined and evaluated in this report. They differ on the particular implementation and the type of data they have been applied to.

The first two methods use the same type of representation: it is a bag of words gathered from entities connected to each element in the RDF graphs. These are built by traversing the graph from each resources up to a certain level and collecting all textual annotations attached to resources. A statistical translator is used to translate these documents into each other languages or into English. The last method, instead of translating the virtual documents replaces the collected terms by identifiers from a multilingual resource (here BabelNet).

In all cases, resources are described by a bag of elements in the same space (target language or BabelNet). A similarity is computed with classical information retrieval techniques (TF or TF·IDF) and a set of links is extracted from the similarity with different modalities (greedy or optimal).

The two techniques have been evaluated against different data sets:

- Parts of DBPedia in English and XLore in Chinese, for experiments 1 and 3;
- The TheSoz thesauri reduced to resources in French, English and German, for Experiment 2.

These experiments have considered many different parameters. Beside normalizing documents cases and removing stop words, preprocessing has little impact. Globally TF·IDF and the best match extraction method provide the best results.

The translation methods have globally performed better than the multilingual resource based method. One intriguing finding is that in the first experiment, the smallest traversal level gave the best results, though in the second experiment the larger level was best. We conjecture that this is due to the type of considered entities: individuals against generic entities.

The results in this deliverable have been reported in [Lesnikova et al. 2014] (Chapter 3), and [Lesnikova et al. 2015] (Chapter 5).

DOCUMENT INFORMATION

Project number	ANR-NSFC Joint project	**Acronym**	Lindicle
Full Title	跨语言环境中语义链接关键技术研究 Linked data interlinking in a cross-lingual environment Liage des données dans un environnement interlingue		
Project URL	`http://lindicle.inrialpes.fr/`		
Document URL			

Deliverable	**Number**	4.2	**Title**	Algorithms for cross-lingual data interlinking
Work Package	**Number**	4	**Title**	Cross-lingual data interlinking

Date of Delivery	**Contractual**	M30	**Actual**	2015-06-30
Status	final		final ☒	
Nature	prototype ☐ report ☒ dissemination ☐			
Dissemination level	public ☒ consortium ☐			

Authors (Partner)	Tatiana Lesnikova, Jérôme David, Jérôme Euzenat			
Resp. Author	**Name**	Tatiana Lesnikova	**E-mail**	Tatiana.Lesnikova@inria.fr
	Partner	INRIA		

Abstract (for dissemination)	Linked data technologies enable to publish and link structured data on the Web. Although RDF is not about text, many RDF data providers publish their data in their own language. Cross-lingual interlinking consists of discovering links between identical resources across data sets in different languages. In this report, we present a general framework for interlinking resources in different languages based on associating a specific representation to each resource and computing a similarity between these representations. We describe and evaluate three methods using this approach: the two first methods are based on gathering virtual documents and translating them and the latter one represent them as bags of identifiers from a multilingual resource (BabelNet).
Keywords	data interlinking, cross-lingual link discovery, owl:sameAs

Version Log			
Issue Date	**Rev No.**	**Author**	**Change**
20/04/2015	1	J. Euzenat	Template
22/06/2015	2	T. Lesnikova	Added chapters for MT and BabelNet
23/06/2015	3	T. Lesnikova	Refactoring of method from chapters
25/06/2015	4	T. Lesnikova	Added introduction
25/06/2015	5	J. Euzenat	Added biblio
26/06/2015	6	T. Lesnikova	Modified introductions, fixed all references
29/06/2015	7	T. Lesnikova	Fixed UTF-8
29/06/2015	8	J. Euzenat	Added extended abstract

Table of Contents

1. Introduction

Linked Data enables the extension of the Web based on Semantic Web technologies. The Semantic Web provides technologies such as the Resource Description Framework (RDF) [Lassila and Swick 1999] for representing data on the web. RDF (Resource Description Framework) is a W3C data model according to which a resource is described by triples (subject, predicate, object). The RDF statements form a directed labeled graph where the graph nodes represent resources and edges represent relations between these resources. A set of statements about a resource constitutes a description set which contains certain characteristics of a resource and thus can be viewed as a ground for a resource "identity".

Knowledge can be expressed in different languages. The project of DBpedia[1] provides a semantic representation of Wikipedia in which multiple language labels are attached to the individual concepts, and has become the nucleus for the Web of Data. Though there are interlingual links between different language versions of Wikipedia, there are knowledge bases in other languages which are not interlinked. For example, XLore [Wang et al. 2013] is an RDF Chinese knowledge base which provides a semantic representation of national knowledge sources (Baidu baike, Hudong baike). Other publishers as the French National Library [Simon et al. 2013], the Spanish National Library [Vila-Suero et al. 2012], the National British Museum[2] make their data available using RDF model in their own language.

Cross-lingual interlinking consists in discovering links between entities across knowledge bases of different languages, see Figure 1.1. It is particularly difficult due to several reasons: (1) the structure of graphs can be different and the structure-based techniques will not be much of help; (2) even if the structures are similar to one another, the properties themselves and their values are expressed in different natural languages. In this regard, we adopt a Natural Language Processing (NLP) approach to address the problem of finding the same object described in two different languages. Our hypothesis is that if two resources denote the same real-world object, then the descriptions of these resources should overlap with each other.

Problem description. Given two RDF data sets with resources described in different languages, the same entity represented in different data sets has to be identified. At the instance level, the values of properties are in different languages, which makes it harder to merge data about the same entity from different sources.

The goal of our research is to identify the same entities across multilingual RDF data sets and link them by owl:sameAs links. For this purpose, we are developing an approach which represents RDF entities as text documents and then compare them. We apply standard Natural Language Processing (NLP) techniques (document preprocessing, term weights, similarity measures) on our data. We particularly explore two strategies [Lesnikova 2014]:

- Applying Machine Translation (MT) in cross-lingual RDF data interlinking [Lesnikova et al. 2014];
- Using references to external multilingual resources [Lesnikova et al. 2015]

The report is structured as follows. In Chapter 2, we describe a general framework for interlinking resources described in different languages. The experiments relying on MT are described in Chapters 3 and 4. The experiment involving an external linguistic resource is presented in Chapter 5. Finally, we propose extensions to the proposed approach in Chapter 6.

[1] http://wiki.dbpedia.org
[2] http://collection.britishmuseum.org/

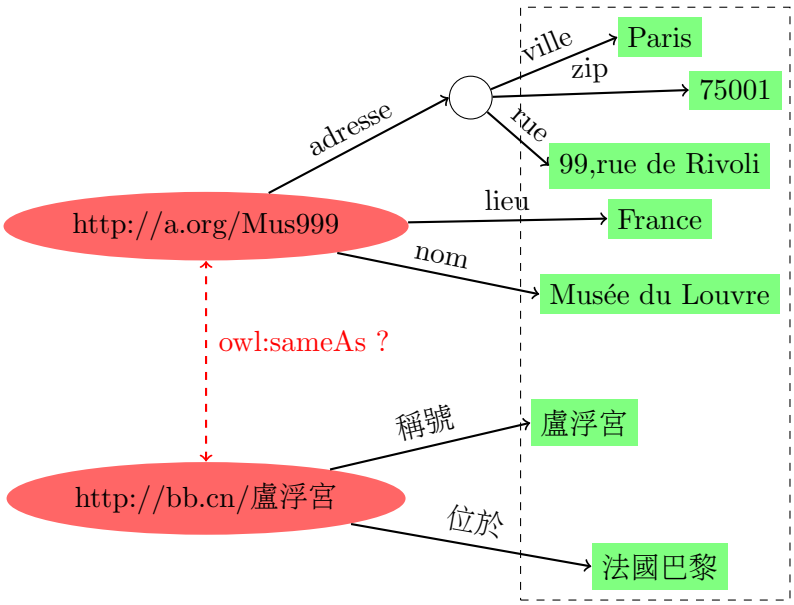


Figure 1.1: Interlinking RDF resources described in different natural languages.

2. General Framework for Cross-lingual RDF Data Interlinking

In this chapter we propose a general framework for cross-lingual interlinking of RDF data. We assume that the resources published in RDF are described in natural languages: property names and literals are usually natural language words. We hypothesize that NLP techniques can be used in order to detect the identical resources and interlink them. This means that our method is designed for RDF data sets which contain descriptions in natural language. It is inappropriate for RDF data sets containing purely numerical values. A set of RDF statements form a labeled directed graph where nodes represent resources and edges represent relationships between these resources. An RDF data set is a graph where resources (individuals or concepts) have labels in natural languages. A context of a resource would be the labels of the neighboring nodes. However a context can be very narrow if there is not much textual information in the description of a concept. Since the method we adopt relies on text similarity, the textual information of a resource is very important: the similarity score highly depends on the overlapping context.

The framework that we designed for interlinking cross-lingual RDF resources is depicted in Figure 2.1.

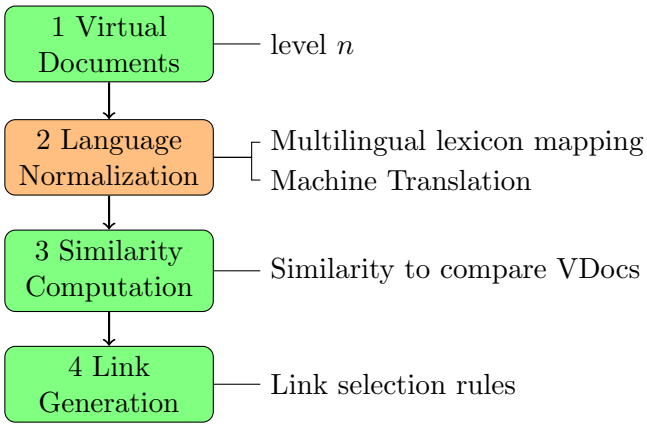


Figure 2.1: Framework for Cross-lingual RDF Interlinking

Given two RDF data sets, the method proceeds as follows.

2.1 Constructing Virtual Documents

First, the resources are represented as Virtual Documents in different natural languages. The notion of a virtual document for RDF resources has been described in [Qu et al. 2006; Lesnikova et al. 2014].

We extract all the language information of a particular resource, for example, such properties as "rdfs:label" and "rdfs:comment" usually contain textual data. The triples of an RDF graph can have simple strings (literals) as an object which serve as a descriptor for a subject. In the example of Figure 2.2, the subject is "dbpedia:Lucerne" which has a label "Lucerne".

The purpose of this extraction is to form a virtual document which contains n levels of language information depending on the specified distance of graph traversal, see Figure 2.2. The language elements attached to a particular type of relationships are taken into account. The property names are not considered. If the object is a literal, it is stored into a virtual document. If not, the algorithm proceeds to the following URI until it collects all the literals

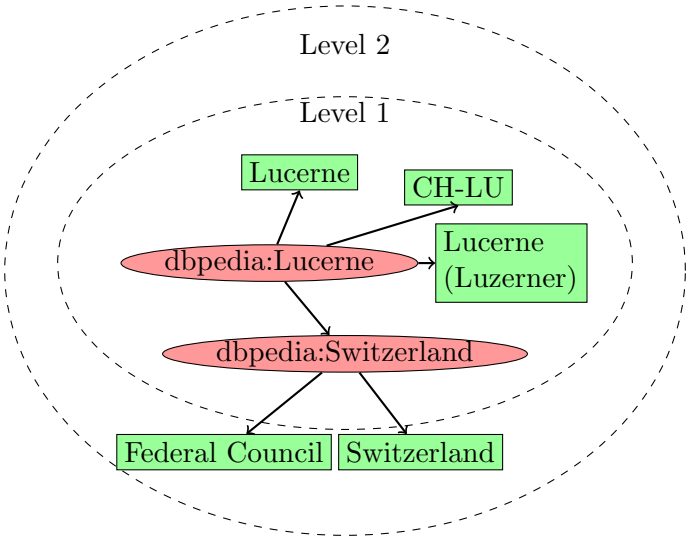


Figure 2.2: Creation of Virtual Documents by Levels

within a given distance. The performance of the method may depend on the amount of text and discriminative power of labels.

2.2 Translating using Machine Translation or Mapping to Multilingual Lexicon

Next, to make these documents comparable we use Machine Translation. Given two virtual documents in two different languages, it is important to make them comparable using a machine translation system. There are different kinds of MT: rule-based, statistical, hybrid. There are Google and Bing translator APIs. At this step, virtual documents in one language can be translated into the other language and vice versa or both languages can be translated into some third language. An alternative approach would be to use Multilingual Lexicon Mapping instead of translation. The terms of virtual documents of both input languages are mapped to common identifiers in a lexicon.

2.3 Preprocessing of Textual Data

Once the terms are translated or replaced by the identifiers, the documents undergo Data preprocessing. Comparable virtual documents are treated as "bags of words", and different number of standard NLP preprocessing techniques (tokenization, stop word removal, etc.) are performed at this stage.

2.4 Computing Similarity

At Similarity Computation stage, various weighting schemes can be used for selecting the discriminant words, for instance, Term Frequency (TF) and Term Frequency*Inverse Document Frequency (TF*IDF) and a similarity method to be applied, for example, the cosine similarity. The output of this stage is a similarity matrix.

2.5 Generating Links

At Link Generation stage, the algorithm extracts links from the similarity matrix. The goal of interlinking is to identify a set of correspondences between concepts. At this stage, an algorithm extracts links on the basis of the similarity between documents. There are different methods to extract alignments. A broad overview is given in [Euzenat and Shvaiko 2013].

In the chapters that follow, we present a series of experiments on interlinking RDF resources of different type (Named Entities or thesaurus concepts) and expressed in different natural languages.

3. Linking Named Entities using Machine Translation

In this chapter, we evaluate the suitability of a Machine Translation approach to interlink RDF resources described in English and Chinese languages. We represent resources as text documents, and a similarity between documents is taken for similarity between resources. Documents are represented as vectors using two weighting schemes, then cosine similarity is computed. The experiment demonstrates that TF*IDF with a minimum amount of preprocessing steps can bring high results.

In this chapter, we describe an experiment on interlinking resources with English and Chinese labels across two data sets. Given two RDF data sets, the goal is to find resources describing the same entity and set an owl:sameAs link between them.

We address the following questions:

- Can Machine Translation and the classical Information Retrieval (IR) vector-space model be suitable for interlinking RDF data?
- How does the quality of generated owl:sameAs links depend on the data preprocessing techniques?

3.1 Translation-based Interlinking Method

The entire data flow with modifiable parameters is illustrated in Figure 3.1.

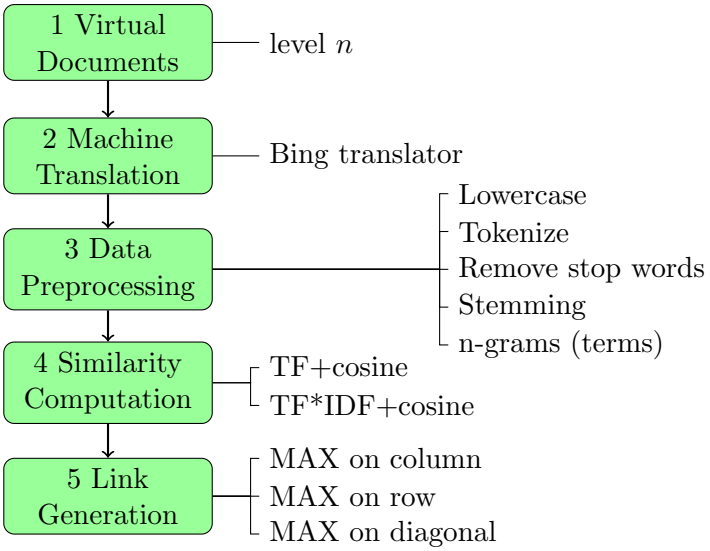


Figure 3.1: Data Flow for Resource Interlinking

Given two RDF data sets, we proceeded as follows.

First, the resources are represented as **Virtual Documents** in different natural languages. To obtain these virtual documents per resource, we collect literals according to the specified graph traversal distance, as described in section 2.1 of Chapter 2.

Next, to make these documents comparable we use **Machine Translation**. Once translated, the documents undergo **Data preprocessing**. We constructed four pipelines so that the number of processing steps is growing with each pipeline.

1. Pipeline 1 = Transform Cases into lower case + Tokenize;
2. Pipeline 2 = Pipeline 1 + Filter stop words;

3. Pipeline 3 = Pipeline 2 + Stem (Porter);
4. Pipeline 4 = Pipeline 3 + Generate n-grams (terms, max length = 2).

In order to compute similarity between the resources, we need to compute similarity between the documents that represent these resources. At **Similarity Computation** stage, we chose two weighting schemes: Term Frequency (TF) and Term Frequency*Inverse Document Frequency (TF*IDF) and applied the cosine similarity. The output of this stage is a similarity matrix. The matrix is such that the virtual documents in the original language are on the vertical axis and the translated documents are on the horizontal axis.

At **Link Generation** stage, the algorithm extracts links from the similarity matrix. We study three ways of extracting links:

1. We select the maximum value in a column only (selecting the best original resource for a translation);
2. We select the maximum value in a row only (selecting the best translation for an original resource);
3. We select the maximum value in a column and a row (selecting such a translation for which the best original document has this translation as best translation).

3.2 Experimental Setup

Our goal is to evaluate how the method described above works and which parameters are important. We also evaluate the suitability of Machine Translation for identifying identical resources.

We would like to observe the effect of the size of virtual documents, preprocessing steps and weighting schemes (TF and TF*IDF) on the results. Basically, we seek an answer to the question: what is the combination of parameters that produces the highest results and can assure the correct match in the interlinking process?

3.2.1 Original RDF Data Sets

The experiment has been conducted on two separate RDF data sets with resources represented in English and Chinese natural languages respectively. Thus, the data consist of the English and Chinese part.

To fulfill the English part, we downloaded the following datasets from DBpedia 3.9[1]: Categories (Labels), Titles, Mapping-based Types, Mapping-based Properties, Short Abstracts, Extended Abstracts. For the Chinese part, we used a part of the Xlore.org[2] data: Abstracts, Reference Links to DBpedia, Inner Links, External Links, Infobox Property, Related Items, Synonyms. XLore is the Chinese knowledge-base Baidu Baike converted into RDF.

All the data files have been accessed via a Jena Fuseki server and its built-in TDB store[3]. Statistics of data loaded into triple stores is presented in Table 3.1.

3.2.2 Test RDF subset

We restricted our experiment to five entity types: Actors, Presidents, US Presidents, Sportsmen, and Geographical places. This was done for observing the difference in similarity within and across types.

[1] http://wiki.dbpedia.org/Downloads39
[2] http://xlore.org/index.action
[3] http://jena.apache.org/documentation/serving_data/

Table 3.1: Statistics about RDF Datasets

	# of classes	# of instances	# of properties	# of triples in total
DBpedia	435	3,220,000	1377	72,952,881
XLore	N/D	262,311	6280	7,063,975

Table 3.2: Experimental parameters

VDocs 2	Pipelines 4	Translation 1	Weight 2	Similarity 1	Link Extraction 3
Level 1 Level 2	Pipeline 1 Pipeline 2 Pipeline 3 Pipeline 4	Bing: ZH→EN	TF TF*IDF	cosine	MAX on column MAX on row MAX on column and row

The Chinese data has already been linked to the English version of DBpedia and we used a list of owl:sameAs links as our reference link set at the evaluation step. Out of the reference link set provided by XLore, we randomly selected 20 instances per category (Actors, Sportsmen, etc.) for which the two linked resources had text in their properties (more than just rdfs:label). In the US Presidents category, there were only 16 linked instances with text, this was compensated by adding four extra presidents into the category of Presidents.

This provided 100 pairs of entities potentially generating 10,000 links.

3.2.3 Protocol

The evaluation was carried out according to the following protocol:

- Provide the two sets of resources;
- Run a method configuration and collect the links;
- Evaluate links against the reference links through precision and recall.

3.3 Evaluated Configuration

The parameters evaluated are presented in Table 3.2. Thus, 48 settings have been explored in total.

Translate ZH into EN

Once we collected a fixed number of entity pairs for each category in the English and Chinese data sets, we needed to make these entities comparable. For our experiment, we used the statistical translation engine: Bing Translator API[4] to translate Chinese virtual documents from the Chinese Simplified into the English language. Sometimes the large documents could not be translated in their entirety, in this case we left everything as is, taking only the part of text that has been translated. It would be interesting to translate documents from English into Chinese as well but our preprocessing tool does not support Asian languages, so at this point we were dealing only with translations from Chinese into English.

Data Preprocessing and Similarity Computation

The tool used for designing our pipelines was RapidMiner[5]. We were using RapidMiner 5.3.013 with the text processing extension.

[4]http://datamarket.azure.com/dataset/bing/microsofttranslator

[5]http://rapidminer.com/products/rapidminer-studio/

Each data preprocessing step corresponds to a particular operator in RapidMiner. For some operators we can specify parameters. Below you can find the parameters used:

- Tokenize: mode: non-letters (i.e. non-letters serve as separators between tokens. Because of this, all dates are not preserved in documents);
- Filter Stopwords (English): built-in stopword list;
- The type of weighting scheme (TF or TF*IDF) was set for each pipeline;
- For computing similarity, we were using Data to Similarity Data operator with cosine similarity.

Link Generation

The output of the similarity computation is a matrix of compared pairs with a value. The 10,000 (100×100) comparisons were tabled as a similarity matrix for evaluation for each tested method. The matrix is such that the vertical axis represents the English DBpedia entities while the horizontal axis represents entities from the Chinese XLore base.

3.4 Results

The obtained results are displayed in Figures 3.2 and 3.3. They show that with TF*IDF/ Level 1 we are able to identify more than 97% of the identical entities. The comparison of virtual documents was done at two levels. The results across and within categories using TF*IDF show the same pattern: the best accuracy is achieved at Level 1 and the results get worse at Level 2. The results for TF were lower than those of TF*IDF so we do not report them here.

The similarity of resources within categories is presented in Figure 3.4. Black squares are 5 categories. The similarities are highlighted according to their value, and the color intensifies as the value grows:

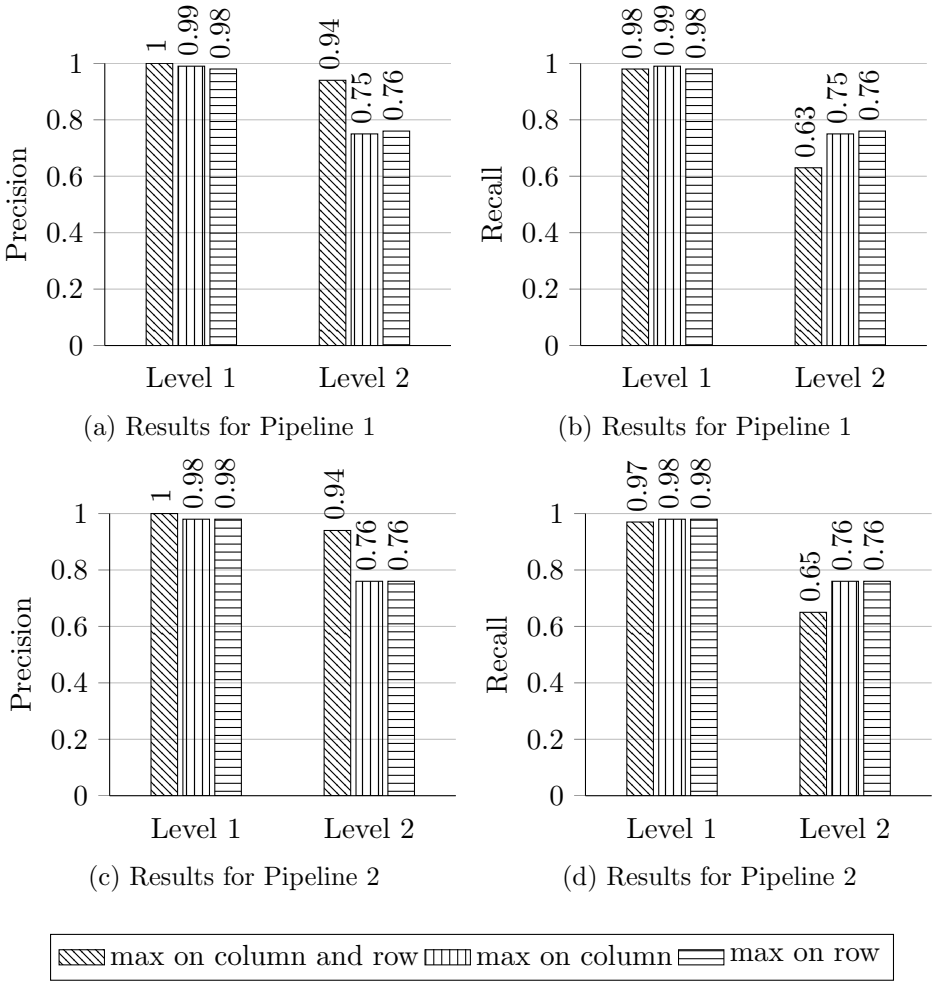
- Values between 0.00 and 0.11 - are suppressed and seen as a white space;
- Values between 0.11 and 0.15 are in light yellow;
- Values between 0.15 and 0.25 are in dark yellow;
- Values between 0.25 and 0.35 are in orange;
- Values between 0.35 and 0.45 are in light red;
- Values between 0.45 and 1 are in dark red.

The correct match is always on the diagonal and the possible confusions are more likely within a category (see the last square (US Presidents)). This is expected since entities of the same type will have much information in common.

3.4.1 Discussion

The main points of the experiment are:

- Our results show the suitability of Machine Translation for interlinking multilingual resources;
- TF*IDF outperforms TF;
- The addition of preprocessing steps seem not to influence the results significantly. The maximum standard deviation is less than 2 points for both precision and recall;
- The quantity of information at Level 1 is usually enough to find a correct match;



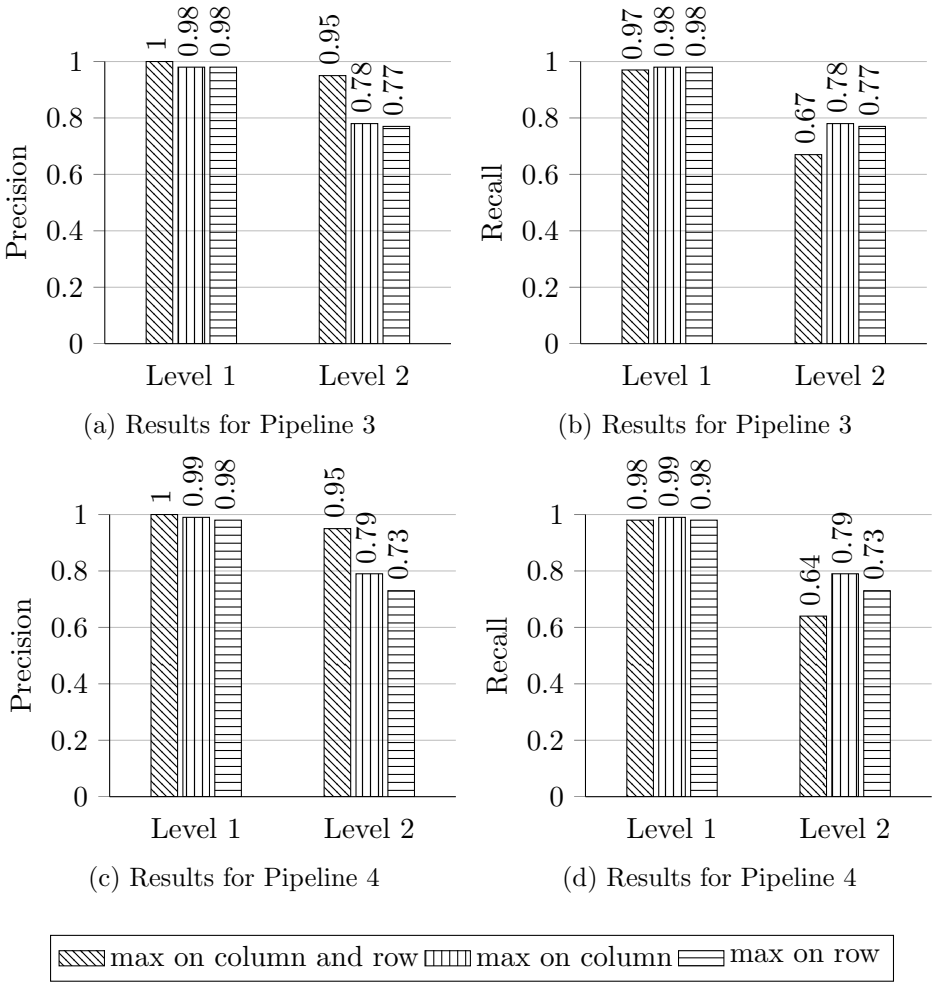
(a) Results for Pipeline 1

(b) Results for Pipeline 1

(c) Results for Pipeline 2

(d) Results for Pipeline 2

Figure 3.2: Results for Level 1 and Level 2 using TF*IDF



(a) Results for Pipeline 3

(b) Results for Pipeline 3

(c) Results for Pipeline 4

(d) Results for Pipeline 4

Figure 3.3: Results for Level 1 and Level 2 using TF*IDF

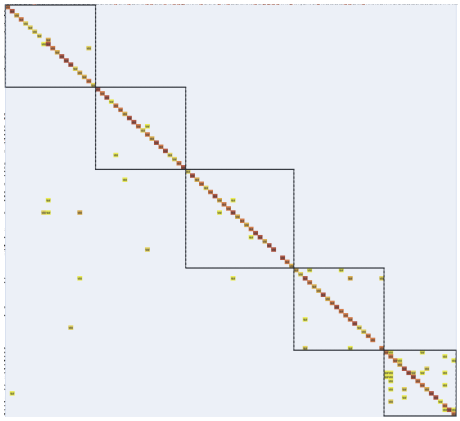


Figure 3.4: Similarity within categories using TF*IDF at Level 1 Pipeline 1. Squares correspond to categories, and the darker the points, the higher the similarity. Dark points on the diagonal are correct matches. Most of the secondary dark points are confined in a square (a single category).

- In general, the results at Level 2 were lower. This may be explained by supposing that the further we go from the node, the more general becomes the information. If there are many shared properties, then at some point many resources will have the same information (this can be due to the structure of the RDF data set). The discriminant information is thus "diluted" and it becomes harder to detect correct correspondences;
- If there is not enough data at Level 1 then by collecting information from Level 2 it is possible to improve the results. This gives us an intuition that the necessity of proceeding to the next level from Level 1 depends on the amount of data at Level 1. We saw this with one of the error cases when comparing across categories.

3.5 Conclusions

Interlinking of resources described in different natural languages across heterogeneous data sources is an important and necessary task in the Semantic Web in order to enhance semantic interoperability. We described an instance-based interlinking method that mostly relies on labels and machine translation technology. The results demonstrated that the method can identify most of the correct matches using minimum information in a resource description with precision over 98%.

Though the reported results provide evidence that our method can be used for finding identical resources across two data sets, there are several axes that we currently left out of scope but will investigate in the future:

- Experimenting with other language pairs;
- Extending the coverage: adding other classes;
- Testing other similarity metrics;
- Exploiting other Machine Translation tools and evaluating their impact on the similarity computation;
- Exploring strategies that do not depend on translation technologies (e.g. mapping to BabelNet).

In the next chapter we describe an evaluation of the translation-based method on the thesaurus concepts.

4. Linking Generic Entities using Machine Translation

Various lexical resources are being published in RDF. To enhance the usability of these resources, related identical resources in different data sets should be linked. If lexical resources are described in different natural languages, then techniques to deal with multilinguality are required for interlinking. In previous work, we designed a method for interlinking Named Entities. At present, we consider how, by using machine translation and taking into account the graph structure of an RDF data set, it is possible to interlink concepts, i.e. generic entities named with a common noun or term. In this paper, we evaluate several methods for interlinking concepts from the TheSoz multilingual thesaurus on three languages: English, French and German. Our results demonstrate that, by taking into account the graph structure of a dataset, we obtain a higher similarity between identical concepts.

In this chapter, we evaluate a translation-based interlinking method on terminology expressed in different natural languages. This interlinking method has been applied to the encyclopedic resources in English (DBpedia [Auer et al. 2007; Bizer et al. 2009]) and Chinese (XLore [Wang et al. 2013]) on which we obtained good results [Lesnikova et al. 2014]. Though this method has been initially developed for interlinking RDF instances with labels expressed in different natural languages, we consider its application to the area such as linking heterogeneous multilingual *linguistic* resources. These lexical-semantic resources are grouped in the Linguistic Linked Open Data cloud[1] [Chiarcos et al. 2011], which is a sub-cloud of the Linked Open Data (LOD) cloud[2]. There are many resources for different languages and domains. Our broad goal is to develop a method that makes no assumption about a particular type of resources as long as these resources are published in RDF.

We address the following problem: Given two thesauri with labels in different languages, find equivalent concepts and link them using owl:sameAs links.

We represent resources as text documents containing labels in a respective language, documents are translated and represented in a vector space model. Similarity between documents is taken for similarity between resources. Extraction of matches is based on the similarity values.

4.1 Translation-based Interlinking Method

The interlinking approach based on machine translation technology has been already proposed in Chapter 2.

The interlinking method consists of five steps:

1. Constructing a **Virtual Document** in different languages per resource, see Figure 2.1. At this step, we suppress all metadata information about the dataset: for example, objects of "http://purl.org/dc/terms/" property describe creators of the dataset, dates of creation and modification. The properties to remove were detected by observing the generated documents. Thus, a virtual document contains only proper lexical items, the names of the properties themselves are also omitted.

2. Translating documents using **Machine Translation** in order to transform documents into the same language.

3. Cleaning documents using **Data preprocessing** techniques. We use the following text preprocessing: Transform Cases into lower case + Tokenize + Filter stop words.

[1] http://linguistics.okfn.org/resources/llod/
[2] http://lod-cloud.net/

4. **Computing Similarity** between documents.
5. **Generating Links** between concepts.

An example of a virtual document at Level 1 before suppressing metadata:

```
stock quotation
4.6.07
```

An example of a virtual document at Level 1 after suppressing metadata:

```
stock quotation
```

An example of a virtual document at Level 2 before suppressing metadata:

```
Descriptor
Descriptors of the TheSoz
...
2011-05-06
2011-05-06
2014-08-14
0.93-en
GESIS - Leibniz-Institut fr
                Sozialwissenschaften
GESIS - Leibniz Institute for the
                Social Sciences
http://www.gesis.org/das-institut/impressum/
http://www.gesis.org/en/institute/impressum/
stock quotation
Finance (e.g. Taxes, Currency)
4.6.07
```

An example of a virtual document at Level 2 after suppressing metadata:

```
stock quotation
stock quotation
Finance (eg Taxes, Currency)
```

4.2 Evaluation Setup

The main objective of this evaluation is to assess the performance of the interlinking method on resources which may not contain Named Entities as their labels.

In this section, we first describe the multilingual data used for experiments. Then we describe the parameters used for evaluating the approach.

4.2.1 Data

In order to conduct the evaluation, publicly available datasets with a significant overlap need to be found. A good alternative is one dataset with labels of the same resource in different

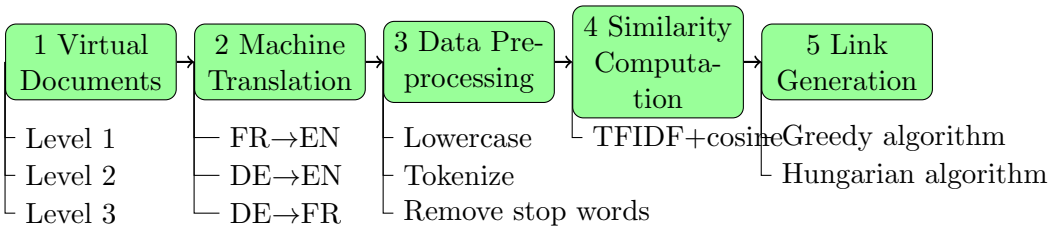


Figure 4.1: Experimental parameters

languages. In this case, one may generate several datasets according to the language of the labels and perform comparison between these newly created datasets.

As a source of a multilingual terminological corpus, we use a multilingual thesaurus for the Social Sciences - TheSoz 0.93 [Zapilko et al. 2013]. This is a SKOS-based thesaurus containing concepts with labels in English, German and French languages. The HTML representation of the thesaurus is available online[3]. Table 4.1 shows information about the concepts in the thesaurus. There are 8223 concepts in total for each language. 12 of them have no English label, and 6 concepts do not have French label. There are 8206 common concepts with a corresponding language label.

Table 4.1: Representation of concepts in each language version of the TheSoz

TheSoz	EN	DE	FR
total number of concepts	8223	8223	8223
concepts without label	12	0	6
number of common concepts	8206	8206	8206

In order to provide a reference alignment, we split the thesaurus into three language specific datasets which contain the same concepts with a label in a respective language. Since the same URI identifies a given concept in each language, we could compare the obtained links against the reference. The dataset consists of 223,574 triples in each language version. In the experiments, we use 8206 concepts shared by three languages.

4.2.2 Evaluated Configuration

The parameters evaluated are presented in Figure 4.1.

Virtual Documents. We constructed virtual documents for Level 1 and Level 2 for the three language pairs. After the results were obtained, we decided to build virtual documents at Level 3 for the best language pair in order to see whether a larger context affects the results.

Translate French and German languages into English. Once we collected virtual documents from the English and French/German data sets, we needed to make these documents comparable. For our experiment, we used the statistical translation system: Bing Translator[4] to translate French and German virtual documents into the English language. Thus, if we compare French virtual documents with the German ones, English is a pivot language.

[3] http://lod.gesis.org/pubby/page/thesoz/
[4] http://datamarket.azure.com/dataset/bing/microsofttranslator

Translate German language into French. In order to verify that the way the virtual documents are translated can affect the results, we also translate German into French, and compare the translated documents against the original French dataset. In this case, the translation is done directly from the source language (DE) into the target one (FR).

Data Preprocessing and Similarity Computation. The tool used for document preprocessing was RapidMiner[5]. We were using RapidMiner 5.3.013 with the text processing extension. Each data preprocessing step corresponds to a particular operator in RapidMiner. For some operators we can specify parameters. The parameters used were:

- Tokenize: mode: non-letters;
- Filter Stopwords (English, French): built-in stopword lists;
- The TF*IDF weighting scheme was used in all settings;
- For computing similarity, we were using Data to Similarity Data operator with cosine similarity.

We assume that the greater the overlap in content (description of an entity), the higher the chance that terms have the same meaning and level of specificity. The discriminability of a term for a given entity will depend on the frequency of this term in a collection of documents.

Link Generation. The output of the similarity computation is a matrix of similarity values between compared entity pairs. We use the Hungarian [Munkres 1957] and greedy algorithms to extract the match assignments. The Hungarian algorithm yields the global optimum: a complete matrix is an input for this algorithm. While the greedy algorithm yields a local optimum: we suppressed null similarities before match extraction.

Additional experiment with randomly removed concepts. The original 8206 concepts common to three language-specific datasets are in a one-to-one relationship with each other. We conducted an additional experiment in order to see how the similarity behaves if concepts in one dataset do not appear in the other one. This experiment has been done on the language pair which showed the highest results using the evaluated configuration described in 4.2.2: EN-DE language pair. We randomly suppressed 40% of concepts from both datasets and only 60% of the concepts has been preserved. Thus, out of 8206 original concepts, only 4943 concepts took part in the experiment, 2995 out of which constituted reference links.

4.2.3 Protocol

The evaluation was carried out according to the following protocol:

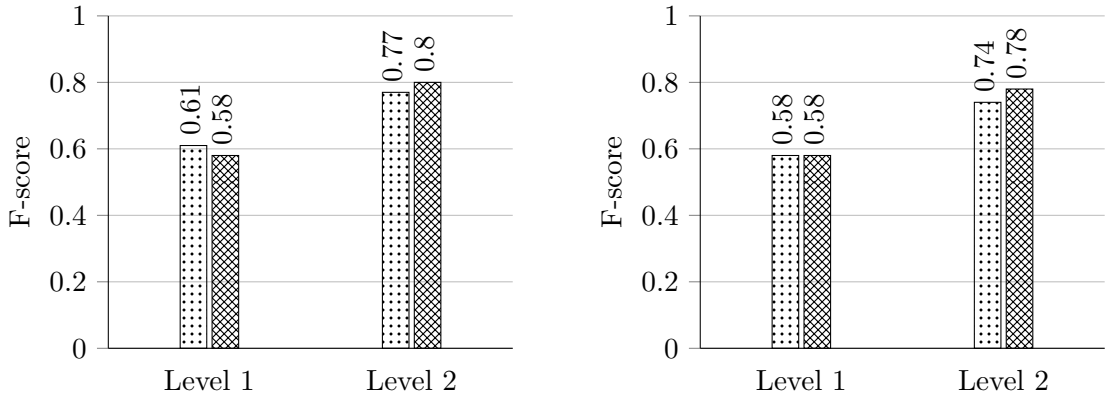
- Provide the two sets of resources;
- Run the method and collect the links;
- Evaluate links against the reference links through precision, recall and F-score.

4.3 Results

The results where French and German virtual documents have been translated into English and compared against the original English data are provided in Figure 4.2. The results

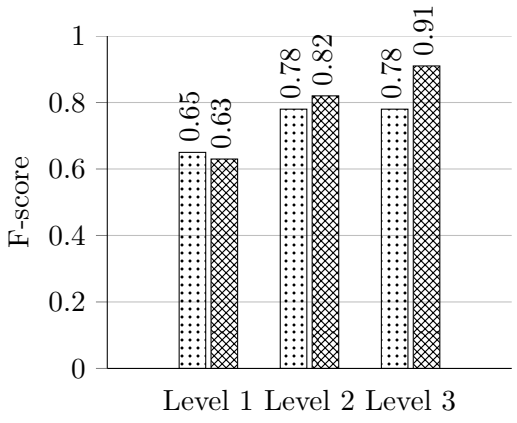
[5]http://rapidminer.com/products/rapidminer-studio/

of comparison against French original data where German virtual documents have been translated into French are presented in Figure 4.3. Finally, Figure 4.4 shows the results of the additional experiment with randomly removed concepts. Each subfigure shows results for a particular language pair using both link extraction algorithms, we compute the F-score for each setting and present it on the y-axis.



(a) Results for the EN-FR language pair

(b) Results for the FR-DE language pair



(c) Results for the EN-DE language pair

Figure 4.2: French and German languages are translated into English and compared against the English original data. For FR-DE pair, English is a pivot language. Results for Level 1, Level 2 and Level 3 using TF*IDF.

In the present experiments, the best F-score of 91% was found at Level 3 which is an improvement by 28 percentage points compared to Level 1. This is totally different from results obtained with Named Entities. In previous experiments [Lesnikova et al. 2014], the cross-lingual interlinking has been done between resources representing Named Entities, and the method could identify most of the correct matches with F-score over 95% at Level 1.

Regarding the impact of machine translation, if machine translation is not exact at Level 1, the mismatch is guaranteed. This is because Level 1 often contains a single word or a short phrase. Given such a string, statistical machine translation returns the most frequent translation, and if it happens to differ with the term in the target language, the result is a failure. That is why it is important to extend the context of a term by proceeding to further levels.

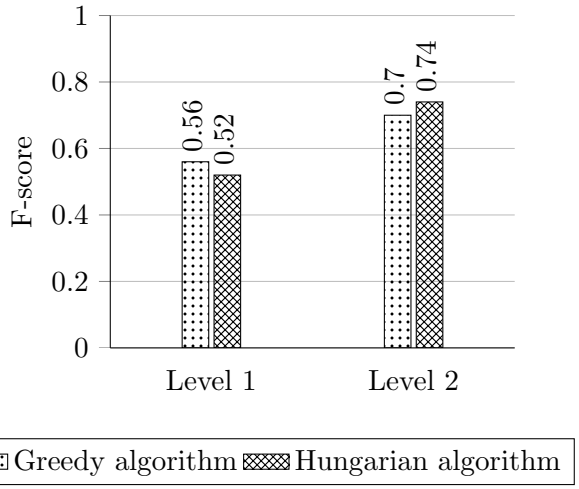


Figure 4.3: Results for the FR-DE language pair. German language is translated into French, comparison done against French original data.

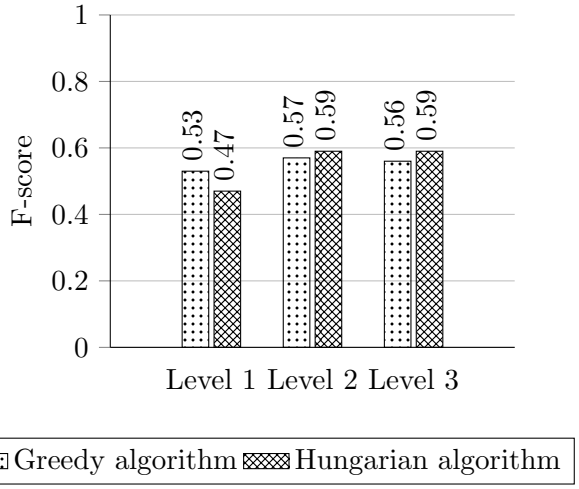


Figure 4.4: Results for the EN-DE language pair. 40% of the concepts have been randomly removed from both datasets.

Concerning the link extraction methods, both link extraction algorithms obtained relatively similar results at Level 1; the Hungarian algorithm outperformed the greedy one at Level 2 and Level 3 and showed an increase of F-score.

The best results are obtained for the English-German language pair (Figure 4.2). The worst results relate to the French-German language pair. This is probably due to the fact that each of these languages has been translated into English and the comparison has been done between the English translations of the terms. Though the difference with other results is not very significant, it may indicate that the meaning was not preserved during translation and scope of the terms in the original versions has been different and thus not preserved after translation. The results of the French-German language pair have not been improved even when the German language has been directly translated into French (Figure 4.3). This may indicate that the machine translation engine is better developed for the German-English language pair.

The results of the experiment with randomly selected concepts show that the similarity

between entities grows as the level increases: precision has been relatively the same across all levels, and we observed an increase of recall by at least 10 percentage points from level 1 to further levels. Though the results are lower, the correct matches have got the highest similarity values even when resources are not in a one-to-one relationship. The best matches are obtained at Level 2 and 3 with F-score of 59% for the Hungarian method.

4.3.1 Errors

We analyzed the errors occurring in the EN-DE language pair (as per results in Figure 4.2) which showed the highest results. A false positive (FP) link is an extracted link which is not in a reference. A false negative (FN) link is a link which is in a reference but was not extracted. Increasing level improves both precision and recall (i.e., both FP and FN decrease). We specifically test if FP and FN decrease monotonically across levels. To that extent we measured:

– the ratio of new FP links introduced when level n increases: $\frac{|FP_{n+1} \backslash FP_n|}{|FP_{n+1}|}$;

– the ratio of new FN links when level n increases: $\frac{|FN_{n+1} \backslash FN_n|}{|FN_{n+1}|}$.

The results are shown in Table 4.2, and we observe that the errors on links are not monotonic.

Table 4.2: Errors made on links

Greedy	L1 → L2	L2 → L3
FP	0.74	0.92
FN	0.08	0.41

Hungarian	L1 → L2	L2 → L3
FP	0.79	0.80
FN	0.05	0.13

Table 4.3: Errors made on entities

Greedy	L1 → L2	L2 → L3
FP	0.34	0.43
FN	0.08	0.41

Hungarian	L1→ L2	L2→ L3
FP	0.05	0.13
FN	0.05	0.13

We hypothesize that the system does not make the same errors but the errors are made on the same entities. So, we measured the ratio of new entities that appear in FP and FN links when level increases. The ratios are the same but instead of links we use entities. The results are shown in Table 4.3. The calculation has been performed on entities which appear in the found links (only unique occurrence of an entity is taken into account (duplicates are removed)). We observe that the errors by the Hungarian method tend to be monotonic and the errors seem to be made on the same entities.

4.3.2 Discussion

The conducted evaluation showed a different performance of the interlinking method when tested on the resources represented by generic terms (a concept label is usually a common noun or a term in a thesaurus). Thus, it seems that it is more difficult to interlink concepts of a thesauri rather than resources corresponding to Named Entities. The findings suggest that the interlinking strategy (including the automatic selection of levels) may depend on the type of entities to be interlinked. Our hypothesis is that comparison of entities belonging to

different types can be done at different information levels. If data is about Named Entities, then it is sufficient to collect information from the resource's closest literals; the further we traverse the graph, the more noise is introduced into the description of the resource (many resources will have similar information and it is harder to find an equivalent entity). If data is about general concepts (as in a thesaurus), then the further we traverse the graph, the information becomes more discriminant and it is easier to find an equivalent entity. We will investigate this hypothesis in the future work.

4.4 Conclusions

In the Semantic Web, RDF data sets may be published with resources labeled in different natural languages. In this context, data interlinking requires specific approaches in order to tackle cross-lingualism. This paper evaluated the efficiency of machine translation when used together with the RDF data structure, as well as the impact of textual information in the description of a resource. We evaluated the approach on 8206 thesaurus concepts in English, French and German languages from the social science domain. We compared the links obtained by two popular assignment algorithms: Hungarian and greedy. In our previous evaluation performed on English-Chinese Named Entities from RDF encyclopedias (DBpedia and XLore), the highest results have been achieved at Level 1 with precision over 0.98. In contrast to those results, the best results have been obtained at Levels 2 and 3. The highest result with F-score of 0.91 has been obtained at Level 3 for the EN-DE language pair. The best correspondences have been extracted by the Hungarian algorithm. The present evaluation shows that the similarity-based method can be applied on resources which do not necessarily contain a Named Entity as their label, though it is harder to find a correct correspondence in this case. We presented a translation-based instance linking approach which can be further tested along several dimensions: (1) evaluation of the method on multilingual ontologies; (2) use of other machine translation engines; (3) evaluation of the method on other language pairs.

The directions for future work may include: (1) context-based matching: matching the French-German DBpedia through the TheSoz mediation and the French-Chinese DBpedia-XLore matching through XLore mediation; (2) using external lexical resources for cross-lingual data interlinking.

In the previous chapters we evaluated the translation-based approach. In the next chapter, we evaluate the lexicon-based interlinking approach.

5. Cross-lingual Linking Using Multilingual Lexicon

In this chapter, we describe an instance interlinking method based on a multilingual lexicon which serves as a pivot language in order to make two resources comparable. We describe an experiment on interlinking resources with English and Chinese labels across data sets and compare it with a translation-based method. Given two RDF data sets, our goal is to find the identical resources and interlink them with owl:sameAs link. We address the following questions:

- Is a multilingual lexicon an appropriate medium to represent documents in two different languages?
- What method performs better: a method based on translation technology or multilingual lexicon?

5.1 Lexicon-based Interlinking Method

The interlinking method is schematized in Figure 5.1.

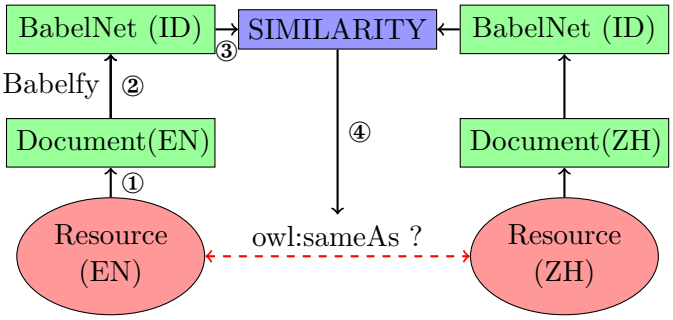


Figure 5.1: Interlinking Method Using Multilingual Lexicon. Multilingual terms are mapped to a common identifier. Similarity is computed between identifiers. Numbers correspond to the steps of the method.

In particular, the method is the following:

1. Constructing a **Virtual Document** per resource.
2. Replacing document terms by identifiers from a **Multilingual Lexicon** in order to project the words of each language onto the same semantic space. At this step, we represent original documents as vectors of identifiers (IDs). A corresponding identifier (ID) is retrieved for a term. An identifier stands for a sense of a term and very often there are many senses (IDs) per term. If more that one sense exists, word sense disambiguation techniques shall be applied in order to select the best sense. The terms not found in a multilingual lexicon are discarded and we do not work with them in our experiments. To compute semantic relatedness, multilingual lexical knowledge resources can be used, e.g., BabelNet [Navigli and Ponzetto 2012] or DBnary [Sérasset and Tchechmedjiev 2014].
3. **Computing Similarity** between documents. We use a standard term weighting scheme (TF*IDF) and apply cosine similarity. These are classical techniques for finding similar documents, moreover, they showed good performance in our previous experiments. The output of this step is a set of similarity values between pairs of virtual documents.

4. **Generating Links** between identical resources. At this stage, an algorithm extracts links on the basis of the similarity between documents. We use the Hungarian or greedy methods to extract links.

5.2 Evaluation Setup

Our goal is to evaluate how the method described above works and what parameters are important. We particularly focus on four parameters: the presence or absence of non-matching entities in a data set, the presence or absence of rdfs:label property value in a virtual document, the amount of text in a virtual document per resource and the link extraction mechanism. We also evaluate the suitability of multilingual lexicon for identifying identical resources.

5.2.1 RDF Data

The experiment has been conducted on two separate RDF data sets with resources represented in English and Chinese respectively. Thus, the data consist of the English and Chinese part. For the English part, we used DBpedia 3.9[1], for the Chinese part – Xlore.org[2]. We restricted our experiment to named entities, e.g, presidents, sportsmen, geographical places. The original data set is the same as described in [Lesnikova et al. 2014], however we have enhanced it in several aspects. The Chinese data has already been linked to the English version of DBpedia and we used a list of owl:sameAs links as our reference link set at the evaluation step. Two datasets have been constructed:

- Original set: contains 100 resources in one-to-one correspondence in English and Chinese languages.
- Original set + noise: we added 10 entities into each language side which do not have a match in the other language. This has been done in order to observe how similarity works when entities do not have matches.

Each of these datasets contains virtual documents of two kinds: with an rdfs:label property value or without it. Thus, we have two variations of each dataset per language: Label and NoLabel.

Since we are linking named entities, an rdfs:label property value is usually a name of an entity which can be highly discriminative. By constructing a virtual document without this property value, we estimate the importance of this element in a resource description.

5.2.2 Experimental parameters

The parameters used for interlinking with a multilingual lexicon are presented in Table 5.1.

Multilingual lexicon mapping. We use BabelNet 2.5.1 which is a multilingual lexicon which connects concepts and named entities in a large network of semantic relations called synsets. Each synset represents a given meaning and contains synonyms which express that meaning in a range of different languages. Since many terms can have several synsets, we also made use of Babelfy 0.9 [3] in order to retrieve the best meaning per term. Babelfy had a limit of 3500 characters for input text, so we had to cut documents at level 2.

[1] http://wiki.dbpedia.org/Downloads39
[2] http://xlore.org/index.action
[3] http://babelfy.org/

Table 5.1: Experimental parameters

Label 2	Data 2	VDocs 2	KB 1	Weight 1	Similarity 1	Link Extraction 2
Label NoLabel	Original set Original set + noise	level 1 level 2	BabelNet + Babelfy: EN→ID ZH→ID	TF*IDF	cosine	Greedy algorithm Hungarian algorithm

Table 5.2: Comparison of MT and BabelNet Methods. Similarity between Entities Using TFIDF. The numbers represent precision (P), recall (R) and F-measure (F) for greedy extraction method.

	Greedy	**Machine Translation**						**BabelNet**					
		level 1			**level 2**			**level 1**			**level 2**		
		P	F	R	P	F	R	P	F	R	P	F	R
Label	**Original set**	1	**1**	1	0.86	0.86	0.86	0.91	**0.88**	0.86	0.68	0.68	0.68
Label	**Original set + noise**	0.9	0.94	0.99	0.74	0.77	0.81	0.75	0.76	0.78	0.63	0.66	0.69
NoLabel	**Original set**	0.87	0.86	0.86	0.79	0.79	0.79	0.77	0.75	0.73	0.67	0.67	0.67
NoLabel	**Original set + noise**	0.76	0.79	0.83	0.70	0.73	0.77	0.65	0.66	0.68	0.60	0.63	0.66

Machine translation. We also apply machine translation on the experimental data. We translate virtual documents using Machine Translation in order to transform documents into the same language. We use Bing Translator[4] to translate Chinese documents into English. Once the documents are translated, we preprocess data to prepare it for similarity computation. Virtual documents are treated as "bags of words", and we use standard NLP preprocessing techniques: transform cases into lower case + tokenize + filter stop words. Once the documents are preprocessed, we apply TF*IDF and cosine similarity.

5.3 Results

In the current evaluation, we have compared the results obtained using both methods: MT-based and BabelNet, see Table 5.2 and 5.3. We have compared the results using two popular assignment algorithms: the Hungarian and greedy. The best results have been achieved by the Hungarian algorithm. The best results are obtained at level 1 on data sets with the rdfs:label property. Results at level 2 decrease for both algorithms: this is because information at level 2 becomes less discriminative and more noisy. Results are also lower when non-matching entities are added. In general, machine translation approach outperformed the approach based on multilingual lexicon. This might be due to the better development of MT capability and unavailability of identifiers for some terms in BabelNet.

[4]https://www.bing.com/translator/

Table 5.3: Comparison of MT and BabelNet Methods. Similarity between Entities Using TFIDF. The numbers represent precision (P), recall (R) and F-measure (F) for the Hungarian extraction method.

	Hungarian	Machine Translation						BabelNet					
		level 1			level 2			level 1			level 2		
		P	F	R	P	F	R	P	F	R	P	F	R
Label	**Original set**	1	**1**	1	0.94	0.94	0.94	0.88	**0.88**	0.88	0.83	0.83	0.83
	Original set + noise	0.9	0.94	0.99	0.83	0.87	0.91	0.73	0.76	0.80	0.7	0.73	0.77
NoLabel	**Original set**	0.93	0.93	0.93	0.92	0.92	0.92	0.81	0.81	0.81	0.78	0.78	0.78
	Original set + noise	0.8	0.84	0.88	0.78	0.82	0.86	0.71	0.74	0.78	0.65	0.68	0.71

5.4 Conclusions

With the growing amount of heterogeneous data on the Web, it is important to make these data machine processable. In the Semantic Web, RDF data sets can be published with labels in different languages. In this context, data interlinking requires specific approaches to tackle cross-lingualism. We have evaluated two approaches based on machine translation and multilingual lexicon. Our results show that the best results are obtained using machine translation with F-measure of 100%, while the results of multilingual lexicon are slightly lower with F-measure of 88%. The highest results have been obtained on datasets with the rdfs:label property which shows that a name of a named entity is a discriminative feature in the interlinking process. Overall, both approaches seem to be promising for cross-lingual RDF data interlinking. However, the limitation would be the availability of language resources for a given pair of languages. The present work can be extended in the following directions:

- Test if both approaches can be complementary: errors made by one method can be corrected by the other method;
- Explore the suitability of Wikipedia for comparing resources.

In the next section we provide several pointers as to how the proposed cross-lingual interlinking framework can be expanded.

6. Conclusion

Multilingual resources (machine translation systems, dictionaries, knowledge-bases, encyclopedias) play an important role in a cross-lingual data interlinking task and are valuable tools for multilingual information processing. Linking entities in a multilingual context relies heavily on such resources. The multilingual resource interlinking can help to uncover the potential of vast amounts of linked open data and facilitate knowledge discovery across language barriers.

In this report we presented a general framework for cross-lingual interlinking of RDF data. The proposed approach relies either on the machine translation technology or multilingual lexicons. The major requirement of the approach is the relatively high quality of machine translation and/or availability of lexical resources per source language. This approach has been evaluated on RDF resources coming from the encyclopedia like DBpedia and thesaurus concepts from the TheSoz. We observed that the machine translation approach showed better results in comparison to the lexicon-based approach. The reasons for that can be a better development of MT technology as well as the disambiguation problems of BabelNet.

6.1 Future work

There proposed framework can be tested further in the following ways:

- Use DBnary as a multilingual lexicon. Dbnary is an effort to provide multilingual lexical data extracted from wiktionary. The extracted data is made available as LLOD (Linguistic Linked Open Data). Linguistic data currently includes Bulgarian, Dutch, English, Finnish, French, German, Greek, Italian, Japanese, Polish, Portuguese, Russian, Serbo-Croat, Spanish, Swedish and Turkish;
- Explore the suitability of Wikipedia for comparing resources: resources are represented as vectors of Wikipedia articles;
- Evaluate the suitability of the framework for ontology matching. Test if XLore ontology in Chinese can be succefully matched to DBpedia ontology in English/French;
- Evaluate cross-lingual interlinking between XLore data and French DBpedia (either by transitivity through the English DBpedia or directly using one of the proposed interlinking methods).

Bibliography

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives (2007). "DBpedia: A Nucleus for a Web of Open Data". In: *6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007*. Vol. 4825. Springer Berlin Heidelberg, pp. 722–735 (cit. on p. 17).

Bizer, Christian, Jens Lehmann, Georgi Kobilarov, S ̈ ̈oren Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann (2009). "DBpedia - A Crystallization Point for the Web of Data". In: *Journal of Web Semantics* 7.3, pp. 154–165 (cit. on p. 17).

Chiarcos, C., S. Hellmann, and S. Nordhoff (2011). "Towards a linguistic linked open data cloud: The Open Linguistics Working Group." In: *TAL* 52(3), pp. 245–275 (cit. on p. 17).

Euzenat, Jérôme and Pavel Shvaiko (2013). *Ontology matching*. en. 2nd. Heidelberg (DE): Springer-Verlag. 520 pp. (cit. on p. 9).

Lassila, Ora and Ralph R. Swick (1999). *Resource Description Framework (RDF) Model and Syntax Specification*. Technical report. World Wide Web Consortium (cit. on p. 5).

Lesnikova, Tatiana (2014). *Interlinking RDF Data in Different Languages*. The TOTh Workshop (Terminology and Ontology : Theories and applications) (cit. on p. 5).

Lesnikova, Tatiana, Jérôme David, and Jérôme Euzenat (2014). "Interlinking English and Chinese RDF Data Sets Using Machine Translation". In: *Proceedings of the 3rd Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (Know@LOD 2014)*. Ed. by Johanna Völker, Heiko Paulheim, Jens Lehmann, Harald Sack, and Vojtech Svátek. Vol. 1243. CEUR-WS (cit. on pp. 2, 5, 7, 17, 21, 26).

— (2015). "Interlinking English and Chinese RDF Data Using BabelNet". In: *DocEng '15*. ACM Symposium on Document Engineering Proceedings (cit. on pp. 2, 5).

Munkres, James (1957). "Algorithms for the Assignment and Transportation Problems". In: *Journal of the Society for Industrial and Applied Mathematics* 5.1, pp. 32–38 (cit. on p. 20).

Navigli, Roberto and Simone Paolo Ponzetto (2012). "BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network". In: *Artificial Intelligence* 193, pp. 217–250 (cit. on p. 25).

Qu, Yuzhong, Wei Hu, and Gong Cheng (2006). "Constructing Virtual Documents for Ontology Matching". In: *Proceedings of the 15th International Conference on World Wide Web*. Edinburgh, Scotland: ACM Press, New York, NY, pp. 23–31 (cit. on p. 7).

Sérasset, Gilles and Andon Tchechmedjiev (2014). "Dbnary: Wiktionary as Linked Data for 12 Language Editions with Enhanced Translation Relations". In: *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing, LREC 2014*, pp. 68–71 (cit. on p. 25).

Simon, Agnès, Romain Wenz, Vincent Michel, and Adrien Di Mascio (2013). "Publishing Bibliographic Records on the Web of Data: Opportunities for the BnF (French National Library)". In: *ESWC*. Ed. by Springer. Vol. Lecture Notes in Computer Science. 7882, pp. 563–577 (cit. on p. 5).

Vila-Suero, Daniel, Boris Villazón-Terrazas, and Asunción Gómez-Pérez (2012). "datos. bne. es: a library linked dataset". In: *Semantic Web Journal* 4.3, pp. 307–313 (cit. on p. 5).

Wang, Zhigang, Juanzi Li, Zhichun Wang, Shuangjie Li, Mingyang Li, Dongsheng Zhang, Yao Shi, Yongbin Liu, Peng Zhang, and Jie Tang (2013). "XLore: A Large-scale English-Chinese Bilingual Knowledge Graph". In: *International Semantic Web Conference (Posters & Demos)*. Vol. 1035. CEUR Workshop Proceeding. CEUR-WS.org, pp. 121–124 (cit. on pp. 5, 17).

Zapilko, Benjamin, Johann Schaible, Philipp Mayr, and Brigitte Mathiak (2013). "TheSoz: A SKOS Representation of the Thesaurus for the Social Sciences". In: *Semantic Web journal (SWJ)* 4(3), pp. 257–263 (cit. on p. 19).